ANALYZING THE FLOW OF INFORMATION FROM INITIAL PUBLISHING TO WIKIPEDIA

An Undergraduate Research Scholars Thesis

by

NATHAN VILLANUEVA

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. James Caverlee

May 2018

Major: Computer Science

TABLE OF CONTENTS

ABSTRACT

Analyzing the Flow of Information from Initial Publishing to Wikipedia

Nathan Villanueva
Department of Computer Science
Texas A&M University

Research Advisor: Dr. James Caverlee
Department of Computer Science
Texas A&M University

This thesis covers my efforts at researching the factors that lead to a research paper being cited by Wikipedia. Wikipedia is one of the most popular websites on the internet for quickly learning about a specific topic. It achieved this by being able to back up its claims with cited sources, many of which are research papers. I wanted to see exactly how those papers were found by Wikipedia's editors when they write the articles. To do this, I gathered thousands of computer science research papers from arXiv.org, as well as a selection of papers that were cited by Wikipedia, so that I could examine those papers and see what made them visible and attractive to the Wikipedia editors.

After I gathered the information on how and when these papers are cited, I ran a series of tests on them to learn as much as I could about what causes a paper to be cited by Wikipedia. I discovered that papers that are cited by Wikipedia tend to be more popular than papers which are not cited by Wikipedia even before they are cited but getting cited by Wikipedia can result in a boost in popularity. Wikipedia editors also tend to choose papers that either showcase a creation of the author(s) or give a general overview on a topic. I also discovered one paper that was likely added to Wikipedia by the author in an attempt at increased visibility.

ACKNOWLEDGEMENTS

I would like to thank my research advisor, Dr. Caverlee, and my honors advisor, Dr. Welch, for their guidance, support, and encouragement throughout the course of this research over the past year. I would also like to thank Texas A&M University and the Undergraduate Research Scholars program for giving me the opportunity to research this material using their resources and for providing aids to assist with writing this thesis.

I would also like to thank TAMU infolab, for all the information, support, and feedback they have provided over the course of this research and for being a great environment to be around others who are doing research in similar areas.

I would finally like to thank the friends that I have gained at Texas A&M, for helping me get through each year and making my college experience as fulfilling as it has been for the past four years.

CHAPTER I

INTRODUCTION

Wikipedia

Wikipedia is currently the fifth most visited website in the world. [1] It achieves this by having a massive repository of informative, concise, noteworthy content. This creation of this content is crowdsourced to everyone on the internet, as indicated by the website's slogan, "The free encyclopedia that anyone can edit". Despite the fact that anyone can edit almost any article they choose, Wikipedia maintains a standard of accuracy comparable to mainstream printed encyclopedias like *Encyclopedia Britannica*. [2]

Wikipedia's popularity makes it a very powerful website. By adding something to an article, an editor can essentially create a "fact" in the public eye for as long as a moderator doesn't see it and take it down. Now, usually when this happens maliciously, a moderator or bot will recognize it as vandalism and remove it almost instantly. The best way for an editor to ensure their addition will not be reversed is to and a source for the information contained within their edit.

Because of this, it is extremely important for people to know where those sources come from and how they tend to be found for use on Wikipedia. The backbone of Wikipedia's accuracy is its extensive use of citations. Every piece of additional information added to Wikipedia must be accompanied by an external citation which backs up the claim. These citations link to all kinds of sources, including news websites, legal documents, and scientific research papers. The last of those is the focus of this research.

Each research paper can be thought of as an "idea" or a "concept" which, once cited by Wikipedia, would be accepted by the general public. If a paper is cited by Wikipedia, the author of that paper has successfully influenced the population in a big way. As a general rule, the more a research paper gets cited, the more likely it is that the information contained in that paper will be seen and taken as fact. If a paper gets cited by something as monolithic as Wikipedia, it is almost certain that the contents of that paper will be spread to a very wide audience.

Possible Issues with Wikipedia

The editors of Wikipedia are not professionals. They have biases, and they have no incentive to contribute to the wiki outside of pure benevolence, so it's very possible they won't put forward the necessary amount of effort required to ensure the articles are 100% accurate. Despite several studies into Wikipedia's accuracy showing that it far more often reliable than it is not, there exist many critics of Wikipedia who worry that its popularity combined with its crowdsourced nature could lead to mass misinformation. [3] While every statement on Wikipedia must be sourced, those sources may not necessarily be reliable for a number of reasons. The most likely problems that could manifest from Wikipedia's crowdsourced methodology are:

- Editor bias, causing one to cite an inaccurate or poorly developed paper that fits their previously drawn conclusions.
- Editor laziness, causing one to pick the first paper they find and accept it as truth without doing more research into the topic.
- Lack of diversity in sources, causing one untrustworthy paper or journal to be greatly overrepresented on Wikipedia as a whole.

- Lack of notability in papers, causing Wikipedia articles to be clogged up with uninteresting, repetitive, or irrelevant information.

My research was based around seeing if the papers cited by Wikipedia could be compromised by any of the above issues.

Existing Research

Due to Wikipedia's massive size and high popularity, lots of research has already been done on the website. "Creating, Destroying, and Restoring Value in Wikipedia", published in 2007, examined the long-lasting edits of Wikipedia's editors, and determined that the vast majority of persistent content on Wikipedia was created by a small minority of users. [4] For the purposes of my research, this meant I could usually consider "Wikipedia editors" as a single entity for simplicity's sake.

Other research on the editors of Wikipedia have revealed that despite the readers of Wikipedia having a 50/50 gender split, 60% of edits are made by men. [5] Another study examined the most controversial topics in the English Wikipedia in 2014 based on the edit wars on their pages. They included George W. Bush, Anarchism and Muhammad. [6]

However, the most interesting and most important studies are about Wikipedia's reliability. One study examined purposeful misinformation on Wikipedia to see how long it generally lasted, and to train up a classifier that could automatically detect and report false information. Another study looked at the completeness of drug information on Wikipedia by comparing it with the information found in textbooks. Both studies had similar results: Wikipedia is pretty accurate overall, but it's far from perfect. The first study found that just under 90% of hoaxes are flagged within an hour, though those other 10% can last for a very long time and pick up a ton of page views. [7] The second study found that around 84% of drug

related information in textbooks could be found in Wikipedia, and that data was around 99.7% accurate. They concluded by saying that Wikipedia was overall "an accurate and comprehensive source of drug-related information for undergraduate medical education". [8]

My research is unique to this research because I am focusing specifically on the sources of Wikipedia, not the text itself. I want to see if the papers the editors use to back up the text they write have any trends or problems inherent to them.

Challenges Faced

Wikipedia's Size

Wikipedia is massive. As of March 13, 2018, Wikipedia has over 5.5 million articles, created by over 33 million registered users submitting over 826 million edits. [9] While there is no information immediately available on how many citations there are on Wikipedia, or how many research papers have been cited by Wikipedia, it is safe to assume there are too large a number to efficiently process. In order to effectively conduct research on Wikipedia and the papers cited by its articles, I would need to find a way to gather a useful sample of Wikipedia articles and papers that had been cited by those articles.

However, while lots of papers have been cited by Wikipedia, substantially more papers have not been cited, and I would need to find a method of sampling those papers as well. Every experiment needs a control group, and I needed a proportional number of papers which had not been cited by Wikipedia in order to create that control group.

Classifying Research Papers

Research Papers have lots of features to analyze. These include several obvious ones such as:

- Title

- Author(s)
- Publisher
- Publication Date
- Area of Study

However, there are also several less obvious, but equally important at the least, features to analyze. These include:

- How and when it has been cited
- Version
- Text used in the abstract
- Text used in the paper itself
- Presence or absence of visual aids
- Scope (does the paper cover a wide range of topics or one very specific topic?)
- Contribution to Field

In designing the analysis portions of my research, I would have to decide which of these features were relevant and which were not. I would also have to figure out a way to either quantify the features I decided were relevant or find specific qualitative categories I could put the papers into for the features where that was not feasible.

Gathering These Features

Many websites exist to function as a database of research papers. Usually these websites will contain the title, author, publishing date, and abstract of the paper. They may or may not have a pdf of the actual paper available. However, many of the features I would need; like the number of times it was cited, the dates it was cited, or a quantified, easily analyzable version of the text; would not be available.

If I wanted to access these features, I would either have to find them at another source and consolidate the information from that source with the information from the original paper database, or I would have to extract the new features from the features I already had.

Finding a Lens

When it finally came time to analyze the papers, I would need a lens to analyze them under. Specifically, what questions were I trying to answer? "Analyzing the papers that get cited by Wikipedia" is an extremely broad scope, and if I wanted to make any headway into actual progress on the subject, I would need to narrow down my efforts into answering one or two substantial questions about the subject.

Our Contributions

This section will provide a brief summary of how I planned to solve the challenges listed in the previous section. I will go into further details in the next two chapters, "Methods" and "Results"

Wikipedia's Size

To get a sample of papers that were cited by Wikipedia, along with a proportional sample of papers that were not cited by Wikipedia, I decided the best course of action was to first gather a large random sample of any research papers, whether they had been cited by Wikipedia or not. I would then separate them based on whether they had been cited by Wikipedia. This would create two proportional samples of papers, one that has been cited by Wikipedia, and one that hasn't.

Classifying Research Papers

When it came time to choose which features were important and which ones were not, I made the following decisions:

- Title – Not Important – I recorded the titles of each paper for record keeping purposes, but when it came to analysis, the titles would not be relevant to whether the paper got cited by Wikipedia or not.
- Author – Important – There is a distinct possibility that Wikipedia editors would be biased towards papers written by certain authors, this makes them worth investigating.
- Publisher – Important – Similar to authors, there is a possibility that Wikipedia editors are biased towards certain journals.
- Publication Date – Important – It is worth checking to see if Wikipedia editors have a recency bias when it comes to choosing papers to cite.
- Area of Study – Not Important – Wikipedia covers all kinds of scientific fields of study. No particular field would reasonably be excluded on the cite as a whole, and when it comes to specific pages, it's pretty obvious that only papers that are relevant to the topic of the page would be included.
- How and when it has been cited – Important – The popularity of a paper could reasonably and significantly impact whether it gets cited by Wikipedia or not. It is also worth investigating if being cited on Wikipedia has any effect on how often it gets cited by other papers after the fact.
- Version – Not Important – How many versions a paper goes through is not usually heavily advertised, and is therefore highly unlikely to affect an editor's ability to find it or their decision to cite it
- Text used in the abstract – Important – The abstract is the first thing an editor would read and could have a significant impact on if they choose to cite it.

- Text used in the paper itself – Important – Similar to above, if the paper is written poorly, it could have a significant negative impact on whether an editor chooses to cite it.
- Presence or absence of visual aids – Not Important – The vast majority of scientific papers in general have some form of visual aid in them, and therefore it goes without saying that the vast majority of papers cited by Wikipedia would have visual aids.
- Scope – Important – It is worth investigating whether Wikipedia editors prefer papers that provide a general overview of a topic, or a deep dive into one very specific issue, or something in between.
- Contribution to Field – Important – What do the papers cited by Wikipedia tend to accomplish? Do they demonstrate a new creation of the authors'? Do they persuade the reader to think about an issue differently? Do they provide a summary of research that has already been made? This is worth investigating.

Gathering These Features

I chose arXiv as an initial database to gather my sample of research papers because of its easy to use API and large collection of scientific papers. For the features that were not available on arXiv, I would either combine or modify the features I had into new features, or I would search for the paper on other websites, like Google Scholar or Scopus, and find the features there.

Finding A Lens

In guiding my research, I decided to answer two primary questions:

What Kind of Papers Get Cited by Wikipedia?

To see if there were any potential problems with the sources that get cited by Wikipedia, I decided to investigate what the general features of papers that got cited were. The features I

looked at included the scope of the paper, the contributions of the paper, when and where the paper was published, the kind of language it uses both in the abstract and the text, and the author. After recording these features for papers that had both been cited and had not been cited by Wikipedia, I would compare the two groups to see if any strong correlations appeared.

How do Wikipedia Editors Find These Papers?

I also wanted to learn more about how Wikipedia editors might find research papers to cite. To do this, I investigated Google Scholar, one of the most popular sites to find research papers, to discover both how easy it was to find the cited papers, and whether these papers had been cited by other papers before they were cited on Wikipedia, which would raise their rankings on Google Scholar. I also investigated the possibility of self-promotion on Wikipedia, how often papers are cited in multiple articles, and how being cited by Wikipedia affects now often a paper gets cited by other papers.

CHAPTER II

METHODS

Data Collection

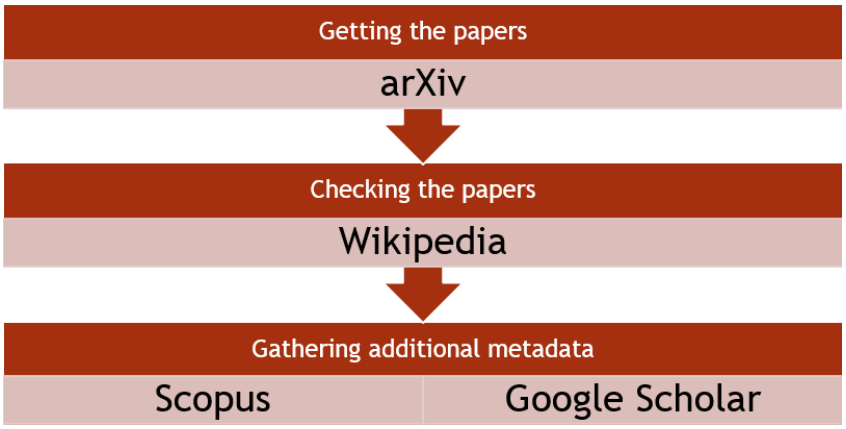

Figure 1. An overview of data flow in my data collection process.

Figure 1 above is a visual representation of a very general overview for how data flowed in my research. All data used in this research was gathered from the internet; either automatically, using python scripts; or manually, using a web browser. Data collected includes research papers and metadata about those papers from arXiv, articles from Wikipedia, and citation information from Google Scholar and Scopus. I first gathered 3000 papers from arXiv, all of which I checked to see if they were cited by Wikipedia. If they were, I gathered additional information on when they were cited from Google Scholar and Scopus. In this section, I will explain how I gathered data from these sources, either manually or automatically.

arXiv

arxiv.org has an easy to use API, allowing easy access to its archived papers with a single "get" call from python's requests library. Each time a request is made to the arXiv API, it returns

an xml file containing information on each paper that matches the search parameters provided. After each call, I would iterate through the list of papers, and for each one record its:

- arXiv ID
- Title
- Author(s)
- Date published
- Journal it was published in (if any)
- Abstract
- Link to pdf

An example of this can be found in Table 1 below.

Table 1. Example Paper

1. arXiv ID	1311.4057v1
2. Title	A Fast Algorithm for Computing High-dimensional Risk Parity Portfolios
3. Author(s)	Théophile Griveau-Billion, Jean-Charles Richard, Thierry Roncalli
4. Date Published	2013-11-16
5. Journal Published In	Journal of Artificial Intelligence Research, Vol 4, (1996), 1-18

6. Abstract	In this paper we propose a cyclical coordinate descent (CCD) algorithm for solving high dimensional risk parity problems. We show that this algorithm converges and is very fast even with large covariance matrices (n > 500). Comparison with existing algorithms also shows that it is one of the most efficient algorithms.
7. PDF Link	https://arxiv.org/pdf/1311.4057.pdf

The second through sixth of those I would save as recorded in a folder for the paper metadata, in a file named [arXiv ID].json. The pdf would be downloaded and saved in a second folder as [arXiv ID].pdf, and then I would extract the text from the pdf and save that in a third folder as [arXiv ID].txt.

Wikipedia

I gathered information from Wikipedia using two different methods, for two different purposes.

The First Method

To determine the number of papers and kinds of papers that are generally chosen by Wikipedia's editors, I took the papers I had already gathered from arXiv and searched to see which of them had been cited by Wikipedia. I used a python script to accomplish this, which used the following procedure. First, it searched the full title of the paper (in quotes) to find any

Wikipedia articles that contained it. This would include any articles that had cited this paper, alongside some false positives.

To differentiate them, the script, using the BeatuifulSoup4 library, would search through each article and look for a journal citation containing both the article's title and its authors. If the page did not contain that, then the article was a false positive in the initial search results. Once the citation was found, then it would be confirmed that the paper was indeed cited by the article.

At that point, the script would search through the article's edit history to find the very first instance of the paper in question being cited. Once the script had that date, it would save the article name and edit date, and move on to the next article in the search results. All of these results would be saved in a folder strictly for papers cited on Wikipedia. Files within that folder would be titled with the arXiv ID, and contain all articles that cite the paper, and the dates when the papers were cited for each page. I collected 3000 papers with this method, of which 30 were cited by Wikipedia. A list of these 30 papers can be found in the Appendix.

The Second Method

Due to the small percentage of research papers that get cited by Wikipedia, the first method did not result in enough papers for me to perform certain kind of analysis. To fix that, I devised a new method of collecting random research papers from Wikipedia, which would be independent of having a proportional amount of papers which were not cited by Wikipedia.

I started on the category page "Areas of Computer Science", since the papers I had gathered previously had been related to the field of computer science in some fashion. This page contains a large amount of links to Wikipedia articles and sub-categories. For the sake of my random collection procedure, this page is the head node in a tree whose children contain the Wikipedia articles (leaves) or sub-categories linked.

For every loop of the script, the procedure will traverse randomly down the tree until it hits a leaf node. Upon reaching said Wikipedia article, after checking to make sure it hasn't been used already, it will search the article for citations that contain one of two keywords: "arXiv:" or "doi:". If it finds a citation with either of these, then it will search for it on arXiv and, assuming it is found, collect the metadata and text for that paper, storing it in a separate folder, so it is not confused with the papers found via the first method.

In both cases, after recording the paper's title and authors, the process will execute a version of the first method, modified to ignore the article we originally found the paper in, on each of these papers, to determine where else they were cited on Wikipedia, but not when, as that is extremely time consuming, even automatically.

Scopus

Scopus, like arXiv, is a research paper database. The primary difference between the two is that while Scopus does not have as many papers as arXiv, it does have information on how and when these papers are cited. Scopus has a usable API, but the relatively small number of papers I had to research and the speed I could access Scopus manually meant that it was faster for me to gather the information manually rather than wait for my request for an API key to get approved.

To get the information on when the papers were cited, I would look the papers up on Scopus based on their title, (sorting by relevance, since the website defaults to sorting by date for some reason) and once I found the paper, usually listed first, I would click the number in the "Cited By" column to get a list of papers which had cited the paper I was looking for.

That's when another feature of Scopus becomes extremely useful. On the page containing the search results is a button labelled "Analyze Search Results." Clicking that button takes you to a page with a list of all the years the paper was cited, as well as the number of times the paper

was cited in each year. The page also contains a line graph visualization of the number of citations per year over time, as can be seen in Figure 2.

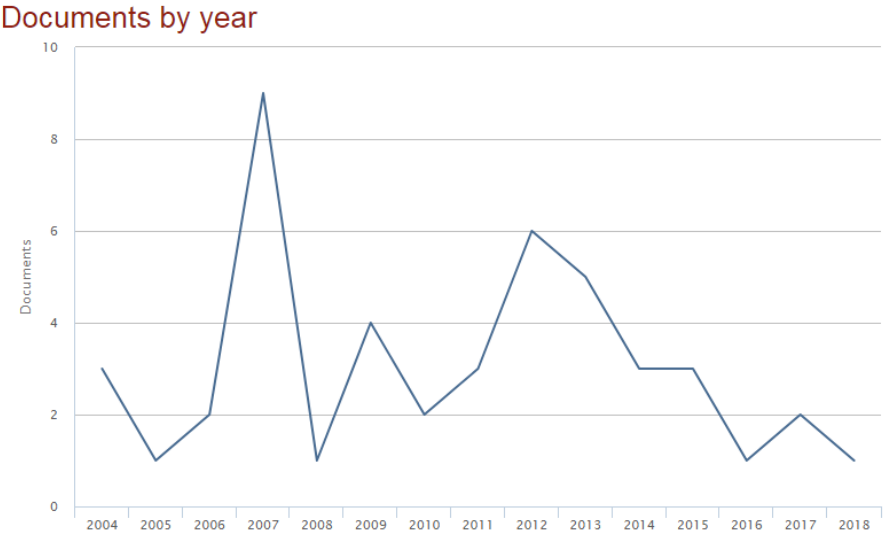

Figure 2. Documents which cite "Building Better Nurse Scheduling Algorithms" in each year.

The numerical data which formed the graph was easily copied and pasted into a text document which was later used for analysis.

Google Scholar

Like Scopus, Google Scholar also contains valuable information on which papers have been cited by which papers. Unfortunately, they do not have a simple centralized location where I can copy numerical citation data from, and they do not allow bots to scrape their website. Because of this, I had to collect any information I needed from them manually, and as a result I only used Google Scholar when the paper could not be found on Scopus.

With Google Scholar, after finding a search result for a paper, one can quickly see any paper that has cited the target paper by simply clicking the “Cited by” option in the search result. For each of my target papers, I looked up all the papers that had cited the target paper, and made a note of their publishing date, as that was the only information that was necessary for the analysis I planned.

Data Cleaning

Most of the data I received from the aforementioned sources was perfectly usable in its current state, but "most" is not "all". For all but one of the sources I collected data from, whether I collected it manually or automatically, I needed to clean the data in order to easily process it. This is how I did that.

arXiv

The arXiv data, for the most part, was relatively easy to clean. None of the papers could be considered unusable, and thanks to arXiv's API, they were all in a consistent format. The only difficult parts of handling the data were dealing with when certain papers were missing fields and normalizing the journals. For the former, not every paper had been published in a scientific journal, so I put some filler text ("None") for those. I also had to use filler text to replace a few missing abstracts and make a note of those when I was doing text analysis on them, so they wouldn't skew the data. For the latter, the journal names arXiv gave me contained the year in a variety of formats which would muddy the analysis. So, I deleted every space separated word that contained a number, which would also remove things like "5th". Fortunately, none of the papers were missing anything critical like a title or a publishing date, so I was able to keep and use all the papers I found.

Wikipedia

No data cleaning was necessary for the Wikipedia data, since the data collection process cleaned the data automatically in a way. As explained previously, instead of collecting the pages themselves, I collected all the papers that were cited by those pages as well as the times those papers were cited.

Google Scholar and Scopus

I collected the same data from Google Scholar and Scopus, namely, for each paper, when they were cited. However, the methods I used to collect that data resulted in two different formats. The Google Scholar data for each paper was collected as a list of years, where each year appeared in the list the same number of times the paper was cited in that year. The Scopus data was collected as a range of unique years between when the paper was published and 2018, and each year was paired with the number of times the paper was cited that year. For analysis purposes, the Scopus data format was the preferable option to store the data in, so, with the help of a very simple python aggregation script, I converted the data I collected from Google Scholar to the format used by the Scopus data.

Data Analysis

After the data was collected and cleaned, it was time to analyze the data and get some results. I analyzed the data using a large variety of different methods, automatically and manually, quantitatively and qualitatively.

Automated Analysis

Most of the automated analysis was simple. The biggest challenge I had to face regarding the automated analysis was that there were a lot of distinctive features that all had to be examined separately.

Citation Count

To examine the citation counts, I simply looked at the Scopus and Google Scholar data I had collected from the papers that were cited by Wikipedia, and computed the mean, median and range of the number of times each paper was cited.

Citation Timing

To examine any possible trends in when each paper was cited in relation to it being cited by Wikipedia, I looked at the annual data when each paper was cited and compared it to the date it was cited by Wikipedia. I eventually ended up using two methods. The first would result in five numbers for each paper:

- Average annual number of citations from two or more years before being cited by Wikipedia
- Number of citations the year before being cited by Wikipedia
- Number of citations the year it was cited by Wikipedia
- Number of citations the year after being cited by Wikipedia
- Average annual number of citations from two or more years after being cited by Wikipedia.

The second would result in as long of a list as was necessary, for however far back and forwards from the citation date the paper had ever been cited.

I then looked at trends with these lists of numbers to see if being cited by Wikipedia had any impact on when how often it was cited by other papers, and vice-versa.

Journal

After normalizing the journal names, I checked the proportions of the journals I found that were cited by Wikipedia and the journals I found that were not cited to see if any journals were severely overrepresented. (Note: To test this feature I used the extended data set for papers cited by Wikipedia)

Author

Similar to the journals, I checked the proportions of each author and their appearance on papers cited by Wikipedia and papers not cited by Wikipedia, to see if any were severely overrepresented. I used the extended data set for papers cited by Wikipedia here as well.

Publication Data

For each paper, I made a note of the time between the date it was published vs when it was cited. I then made a note of the mean, median, range, and standard deviation of those times.

Wikipedia Article Count

To examine how many pages a paper tended to get cited by, I used a python script and the data I had collected to create a histogram of how many papers got cited by how many Wikipedia articles.

Manual Analysis

Certain features could not be assessed automatically. For those features, I needed to manually look at a sample of the data to turn features that would be very difficult for a computer to determine into numerical data that I could draw a conclusion from.

Scope

I decided to sort the papers into two groups base on the size of their scope. I defined a large scope and a small scope as follows:

- Large Scope – A paper with a large scope was defined as a paper which gave a general overview of a relatively general topic. These topics can range from entire fields of science; like data science, online security, or cloud computing; or important questions and possible answers to the chosen question, like the costs and benefits of different encryption methods. Some of the papers I found in this category include “Quantum

algorithms for algebraic problems", which describes the benefits of quantum computing over traditional computing, and "Astroinformatics: A 21st Century Approach to Astronomy", which provides a summary of the field of Astroinformatics.

- Small Scope – A paper with a small scope was defined as a paper which gave a more in depth look into a more specific topic. These topics generally involved either new solutions to previously existing problems, or an investigation on data that pertained to a specific set of circumstances. Some of the papers I found in this category include "The IRAS 1.2 Jy Survey: Redshift Data", in which the redshift data on various galaxies from 1995 is collected and listed, and "An Even Faster and More Unifying Algorithm for Comparing Trees via Unbalanced Bipartite Matchings", where the title is relatively self-explanatory.

After I defined the categories, I manually looked through the papers and split them based on which category they best fit in. I did this separately for papers cited by Wikipedia and papers not cited by Wikipedia.

Contribution

I also defined two categories to determine the contribution the author(s) of any given paper made to their respective field. These categories are defined as follows:

- Description – Authors of papers in this category took previously existing data, collected it, and drew conclusions from it. They used algorithms and methods that already existed to come to new conclusions about some topic. Examples of papers in this category include "On Prediction Using Variable Order Markov Models", which compares several existing algorithms to see which one performs the best, and "Challenges of Big Data

Analysis", which describes several hurdles to efficient big data analysis and provides the authors' perspectives on those hurdles.

- Creation – Authors of papers in this category designed something new and are using the paper to present their new creation to the world. Note that because of how this category is defined, it is extremely unlikely for a paper to be both "Large Scale" and "Creation". While on occasion someone does create an entirely unique field of science, it is very rare, and the samples I gathered for this research contained no examples of this. Examples of papers which fall into the "Creation" category include "Building Better Nurse Scheduling Algorithms" which details a new innovative method of comparing such algorithms, and "A Parametric Simplex Algorithm for Linear Vector Optimization Problems", in which the mentioned algorithm is presented.

While looking through the papers and sorting them by scope, I also sorted them by contribution, creating four sub-categories from the two categories of Scope and the two categories of Contribution. I then made a note of how many papers fit in each category depending on whether they had been cited by Wikipedia or not and compared the differences by percentage.

CHAPTER III

RESULTS

This chapter will be split into separate parts which analyze a separate feature or group of features. Each section will contain a brief reminder of the data involved, followed by the results from my analysis of said data, initial observations about those results, and conclusions on how those features likely impact Wikipedia. I will first list off the features that had interesting, noteworthy, or important results, before briefly listing off the features that ended up not being shown to have a significant effect on a paper being cited by Wikipedia.

Citation Count

Papers cited by Wikipedia had a very wide range when it came to how much they were cited. My analysis of the sample papers I gathered revealed the following information about the number of times papers cited by Wikipedia were cited by other papers:

- Range: 1-2,928 citations
- Median: 46 citations
- Mean: 234 citations

I also noted how many times each paper was cited by other papers before being cited by Wikipedia. Those results were as follows:

- Range: 0-1,479 citations
- Median: 18 citations
- Mean: 115 citations

It should be obvious that the latter set of numbers would be smaller than the former set, considering the papers I gathered tended to continue getting cited after the were cited by Wikipedia.

Of course, these numbers are meaningless without numbers to compare them to. Other people have already done research on how many times papers tend to get cited in general, whether they get cited by Wikipedia or not, so here is what I found regarding that.

- Range: 0-305,148 citations
- Median: 4 citations
- Mean: 8 citations [10][11]

It is rather obvious from looking at these numbers that papers cited by Wikipedia are significantly more popular than the average paper. This is not very surprising, to say the least, but now the data is there to demonstrate. One possible reason for this is that a paper getting cited more often increases its notability on sites like Google Scholar, which is a very popular website for all kind of researchers, including Wikipedia editors, to use to look for sources.

Wikipedia Article Count

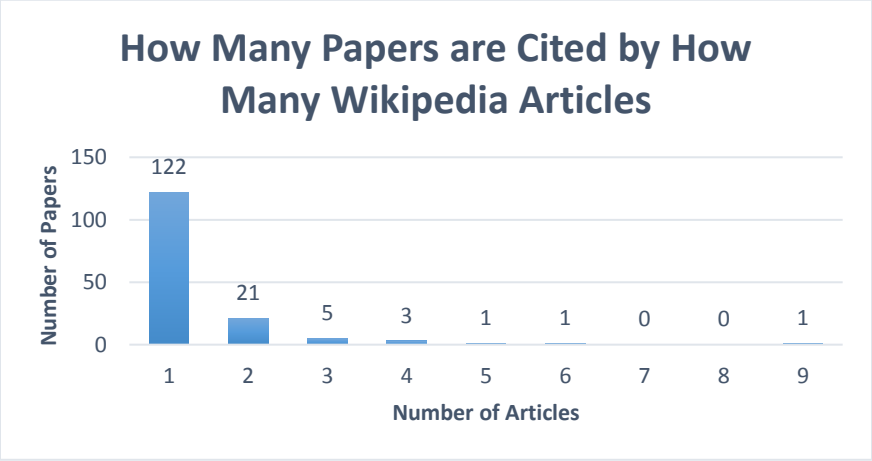

Figure 3. Histogram showing the number of papers that are cited by a certain number of Wikipedia articles.

A popular practice in finding sources is to go to Wikipedia for any needed information, but instead of citing Wikipedia, one would cite the sources. I wanted to see if that practice was popular among the editors of Wikipedia. I called this practice “citation inbreeding”, where editors looking for sources would look within the same place they were planning to use those sources. For Wikipedia editors, this would mean finding sources for an article in the source list for a different article. To see if this was common, I created a histogram showing how many papers were cited by how many articles on Wikipedia. This histogram can be seen in figure 3 above.

The data very clearly shows that inbreeding on Wikipedia is not common, if practiced at all. I was genuinely surprised by this, because Wikipedia articles tend to have a lot of overlap in the material they cover. However, it turns out that the vast majority of articles are cited only once, and it is extremely rare for an article to be cited more than twice.

For curiosity’s sake, I looked at the paper that was cited nine times to see what made it special. The paper that was cited nine times in the histogram was “CODATA Recommended Values of the Fundamental Physical Constants: 2014” released by the Committee on Data for Science and Technology (CODATA) to define what numerical values should be used for important constants like acceleration due to gravity in meters per second squared or atomic mass in grams. It makes sense that this paper is cited as much as it is, since it is extremely general and can be used in a large variety of articles.

Scope and Contribution

I first split papers into two groups, cited by Wikipedia and not cited by Wikipedia. In each of those groups, I split the papers into four groups, Which can be seen in Table 2 and Table 3.

Table 2. Composition of Papers that have been Cited by Wikipedia.

	Small Scope	**Large Scope**
Description	17%	33%
Creation	50%	

Table 3. Composition of Papers that have not been Cited by Wikipedia.

	Small Scope	**Large Scope**
Description	60%	17%
Creation	23%	

As previously mentioned, large scope creation papers are very rare, and none were in my data set.

As shown in the data, while the majority of papers not cited by Wikipedia are small scope description papers, the majority for papers cited by Wikipedia are small scope creation papers. However, large scope description papers are also overrepresented on Wikipedia. This is likely because if an editor is writing a Wikipedia article on a broad subject, they would want to use papers that also provided a broad overview of a subject. Meanwhile, if an editor was writing a paper on a specific subject, rather than get papers that summarize or aggregate the work of others, they could get that work straight from the horse mouth for a more authoritative source.

Journal

Which journal a paper is published in has no obvious effect on whether it gets cited on Wikipedia or not. However, papers that get published in any journals are around 3 times more likely to get cited on Wikipedia than those that are not. Specifically, 1.9% of papers in my data

set that were published in a reputable journal were cited on Wikipedia, while 0.64% of papers that were not published in a reputable journal were cited on Wikipedia. This is fairly reasonable, as papers that were cited in a reputable journal would have higher trust scores on recommender websites like Google Scholar, and would be easier to find.

Citation Timing

As explained earlier, I used two methods to test whether getting cited on Wikipedia had an effect on how much a paper was cited by other papers, or vice versa. It turned out the data from the first method was just a less useful version of the data from the second method, so Figure 4 shows my data from the second method.

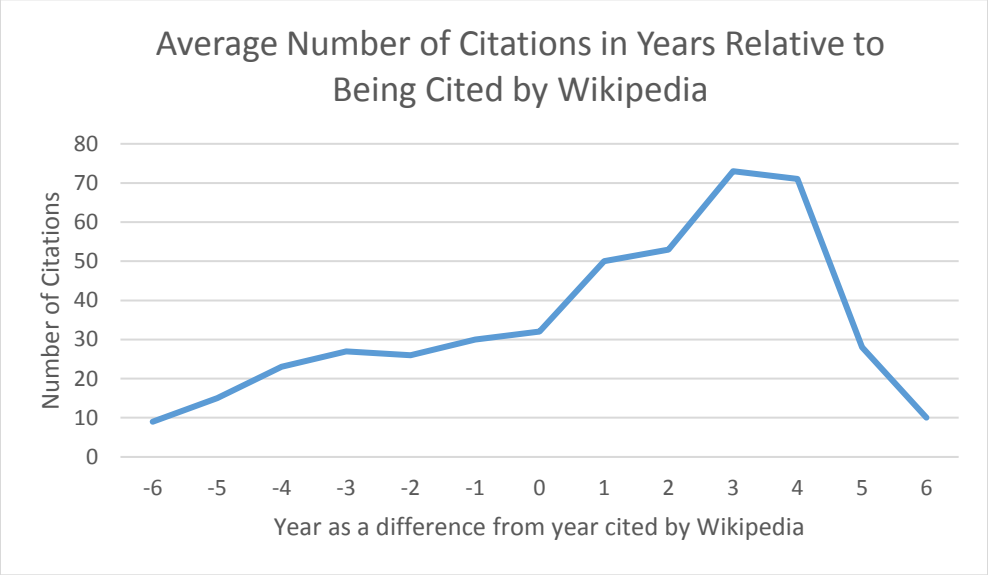

Figure 4. Line graph showing, on average, how many times the papers I collected got cited in the years before and after getting cited by Wikipedia.

While a steady incline can be seen in the years leading up to being cited by Wikipedia, the incline becomes much steeper in the few years after the graph passes "0", the year the paper was cited by Wikipedia. I considered the fact that one paper with a massive number of citations was skewing the average numbers, so I ran the analysis again, but this time I normalized the

numbers by taking the average of the percentages of citations from that year. So if one paper was cited 50 times in a given year, but it was cited 500 times overall, it would count as 10%, not as 50 citations. The results for this are seen in Figure 5 below.

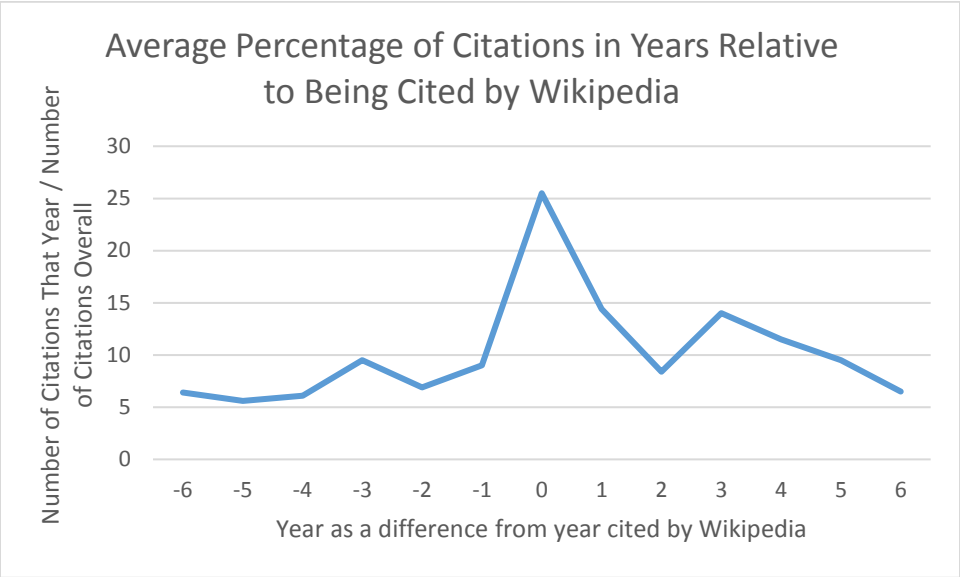

Figure 5. Year difference from year cited on Wikipedia against average citation percentage that year.

While the increase is not as continuous as it was with the raw numbers of citations, There is still a massive increase in citations immediately after a paper gets cited by Wikipedia. This can likely be attributed to the previously mentioned source gathering method of using Wikipedia for information, then citing Wikipedia's sources. If a paper gets cited by Wikipedia, it is likely to get a bump from people using that method.

Self Publication

While testing to see if Wikipedia had a recency bias, I came across a very interesting discovery. Firstly, Wikipedia does not have a recency bias. I looked at all the time differences between a paper being published and a paper being cited on Wikipedia, and the times ranged evenly from two days to over 20 years. However, the paper that was cited on Wikipedia two days

after it was published made me very curious. To put it bluntly, it seemed way too fast for that information to naturally spread to someone who would cite it on Wikipedia.

My hypothesis was that the author of the paper edited Wikipedia himself to make his paper more popular. I had some circumstantial evidence that the paper was not genuine, mainly the extremely brief time between publishing and citing and the fact that after being cited on Wikipedia, it was cited exactly twice, well on the low end of the spectrum for the papers in my data set, but nothing concrete.

However, then I decided to look at the edit history of the article that cited the paper. I found the edit where the paper was first cited in the article, and I looked at the IP address associated with that edit. That address's only other major contribution to Wikipedia was to cite a separate paper that shared an author with the original suspicious paper.

This has significant implications for the integrity of Wikipedia. Everything I had seen so far led me to believe that Wikipedia mainly cited popular and superior quality papers, but in my relatively small sample size, I was able to find a paper that was added to Wikipedia by the author in a blatant conflict of interest. For the record, citing your own original research on Wikipedia is against the site rules. The good news is the methods I used to discover this author's rule breaking habits were entirely quantitative, and that means that there is likely a way to automatically detect and ban this kind of self-promotion.

Non-Significant Features

Very briefly, I'd like to mention the features that I tested, but found no interesting or meaningful results for.

- The author of a paper is insignificant to whether a paper gets cited to Wikipedia or not.

- I ran text classifiers on both the abstracts of papers and the bodies of papers. Neither was able to classify the papers into “Cited” and “Not Cited” groups any better than guessing.
- As previously mentioned, the specific journal a paper is published in and the time between publishing and potential citing by Wikipedia have no effect on whether a paper gets cited or not.

CHAPTER IV

CONCLUSION

Summary of Findings

Over the course of my research I found several interesting and significant facts about Wikipedia and the papers cited thereupon. To summarize:

- Papers cited on Wikipedia, on average, are cited more than papers not cited on Wikipedia. This is still true if you only count the citations that happened before the papers were first cited by Wikipedia.
- The clear majority of papers cited by Wikipedia are only cited on one article. This indicated that while looking at Wikipedia's sources may be a popular method for finding sources among students, it is not popular among Wikipedia editors.
- Papers cited by Wikipedia tend to be either broad overviews of subjects, or authors presenting their creations. There is more of the latter than the former, but significantly more of both of those than overviews of narrow topics.
- Papers cited by Wikipedia experience a bump in citations after being added to the wiki due to the increased exposure.
- This increased exposure can lead some authors to cite their own papers when editing Wikipedia, regardless of the site's rules or the potential conflict of interest.

Limitations

My research, while covering a wide range of topics, was limited by the regretfully small sample size I ended up with. I had 3000 papers randomly chosen from arXiv, but only 30 papers from that random group that were cited by Wikipedia. Because on this, I had to get the remaining

124 papers for my analysis through a semi-random crawl of Wikipedia's computer science section. I was also not as able to explore the idea of self-promotion as much as I wish I was able to. The suspicious paper I found was one in only 30. If I had found it earlier in my research, I would have looked further into that, and possibly developed a method for automatically finding such edits.

Future Research

Ideas for future research would be to basically fix what I explained in the pervious section. Re-running my experiments with a larger sample size would result in far more conclusive results. Another option would be to look at papers that were added and removed from Wikipedia. If one was label to look at the edit history of every page to see when papers were cited and later un-cited, it could result in some very interesting findings.

However, I think the most interesting and practical idea to follow would be to further examine the idea of self-promotion on Wikipedia. If someone was able to create an automatic detector that could stop and reverse when people tried to cite their own paper, it would significantly improve Wikipedia's integrity as a data aggregation tool.

REFERENCES

[1] The top 500 sites on the web. (n.d.). Retrieved September 16, 2017, from http://www.alexa.com/topsites

[2] Giles, J. (2005, December 14). Internet encyclopaedias go head to head. Retrieved January 27, 2018, from https://www.nature.com/articles/438900a

[3] Nielsen, F. A. (2007). Scientific citations in Wikipedia. First Monday, 12(8). doi:10.5210/fm.v12i8.1997

[4] Priedhorsky, R., Chen, J., Lam, S. (., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. Proceedings of the 2007 International ACM Conference on Conference on Supporting Group Work - GROUP 07. doi:10.1145/1316624.1316663

[5] Tancer, B. (2007, April 25). Who's Really Participating in Web 2.0. Retrieved April 09, 2018, from http://content.time.com/time/business/article/0,8599,1614751,00.html

[6] Yasseri, T., Spoerri, A., Graham, M., & Kertesz, J. (2013). The Most Controversial Topics in Wikipedia: A Multilingual and Geographical Analysis. SSRN Electronic Journal. doi:10.2139/ssrn.2269392

[7] Kumar, S., West, R., & Leskovec, J. (2016). Disinformation on the Web. Proceedings of the 25th International Conference on World Wide Web - WWW 16. doi:10.1145/2872427.2883085

[8] Kräenbring, J., Penza, T. M., Gutmann, J., Muehlich, S., Zolk, O., Wojnowski, L., . . . Sarikas, A. (2014). Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology. PLoS ONE, 9(9). doi:10.1371/journal.pone.0106930

[9] https://en.wikipedia.org/wiki/Special:Statistics

[10] Dvorsky, G. (2014, October 30). What Are The Most Cited Research Papers Of All Time? Retrieved March 28, 2018, from https://io9.gizmodo.com/what-are-the-most-cited-research-papers-of-all-time-1652707091

[11] Weingart, S. (n.d.). How many citations does a paper have to get before it's significantly above baseline impact for the field? Retrieved March 28, 2018, from http://www.scottbot.net/HIAL/index.html@p=22108.html

APPENDIX

Table 4. Original 30 Papers Found that were Cited by Wikipedia

Title	Author(s)	Year
Private quantum computation: An introduction to blind quantum computing and related protocols	Joseph F. Fitzsimons	2016
Insecurity of Quantum Secure Computations	Hoi-Kwong Lo	1996
No Signalling and Quantum Key Distribution	Jonathan Barrett, Lucien Hardy, Adrian Kent	2004
Applications of tripled chaotic maps in cryptography	Sohrab Behnia, Afshin Akhshani, Amir Akhavan, Hadi Mahmodi	2007
Coordination in Network Security Games: a Monotone Comparative Statics Approach	Marc Lelarge	2012
The Random Oracle Methodology, Revisited	Ran Canetti, Oded Goldreich, Shai Halevi	2000
Google Android: A State-of-the-Art Review of Security Mechanisms	Asaf Shabtai, Yuval Fledel, Uri Kanonov, Yuval Elovici, Shlomi Dolev	2009
An Authentication Protocol for Future Sensor Networks	Muhammad Bilal, Shin-Gak Kang	2017
Challenges of Big Data Analysis	Jianqing Fan, Fang Han, Han Liu	2013

Table 4 Continued.

Mapping EU fishing activities using ship tracking data	Michele Vespe, Maurizio Gibin, Alfredo Alessandrini, Fabrizio Natale, Fabio Mazzarella, Giacomo C. Osio	2016
Community Structure in Time-Dependent, Multiscale, and Multiplex Networks	Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, Jukka-Pekka Onnela	2009
The Revolution in Astronomy Education: Data Science for the Masses	Kirk D. Borne, Suzanne Jacoby, K. Carney, A. Connolly, T. Eastman, M. J. Raddick, J. A. Tyson, J. Wallin	2009
Evaluating Pricing Strategy Using e-Commerce Data: Evidence and Estimation Challenges	Anindya Ghose, Arun Sundararajan	2006
Automating biomedical data science through tree-based pipeline optimization	Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole A. Lavender, La Creis Kidd, Jason H. Moore	2016
Astroinformatics: A 21st Century Approach to Astronomy	Kirk D. Boone	2009
Adaptive Real Time Imaging Synthesis Telescopes	Melvyn Wright	2012

Table 4 Continued.

The IRAS 1.2 Jy Survey: Redshift Data	Karl Fisher, John Huchra, Michael Strauss, Marc Davis, Amos Yahil, David Schlegel	1995
Providing Authentic Long-term Archival Access to Complex Relational Data	Stephan Heuscher, Stephan Jaermann, Peter Keller-Marxer, Frank Moehle	2004
The NumPy array: a structure for efficient numerical computation	Stefan Van Der Walt, S. Chris Colbert, Gaël Varoquaux	2011
CODATA Recommended Values of the Fundamental Physical Constants: 2014	Peter J. Mohr, David B. Newell, Barry N. Taylor	2015
A Parametric Simplex Algorithm for Linear Vector Optimization Problems	Birgit Rudloff, Firdevs Ulus, Robert Vanderbei	2015
A Fast Algorithm for Computing High-dimensional Risk Parity Portfolios	Théophile Griveau-Billion, Jean-Charles Richard, Thierry Roncalli	2013
On Prediction Using Variable Order Markov Models	Ron Begleiter, Ran El-Yaniv, Golan Yona	2011
On the stability of the Bareiss and related Toeplitz factorization algorithms	Adam W. Bojanczyk, Richard P. Brent, Frank R. de Hoog, Douglas R. Sweet	2010
Quantum algorithms for algebraic problems	Andrew M. Childs, Wim van Dam	2008
A Tutorial on Spectral Clustering	Ulrike von Luxburg	2007

Table 4 Continued.

The Design and Experimental Analysis of Algorithms for Temporal Reasoning	Peter van Beek, Dennis W. Manchak	1996
An Even Faster and More Unifying Algorithm for Comparing Trees via Unbalanced Bipartite Matchings	Ming-Yang Kao, Tak-Wah Lam, Wing-Kin Sung, Hing-Fung Ting	2001
Better algorithms for unfair metrical task systems and applications	Amos Fiat, Manor Mendel	2004
Building Better Nurse Scheduling Algorithms	Uwe Aickelin, Paul White	2008